TOTAL INTERNAL AND EXTERNAL LENGTHS OF THE BOLTHAUSEN-SZNITMAN COALESCENT

GÖTZ KERSTING, JUAN CARLOS PARDO, AND ARNO SIRI-JÉGOUSSE

Abstract. In this paper, we study a weak law of large numbers for the total internal length of the Bolthausen-Szmitman coalescent. As a consequence, we obtain the weak limit law of the centered and rescaled total external length. The latter extends results obtained by Dhersin & Möhle [9]. An application to population genetics dealing with the total number of mutations in the genealogical tree is also given.

1. Introduction and main results

In population genetics, one way to explain disparity is to observe how many genes appear only once in the sample. A gene carried by a single individual is the result of two possible events: either the gene comes from a mutation that appeared in an external branch of the genealogical tree, either this gene is of the ancestral type and mutations occured in the rest of the sample (see Figure 1). We suppose that events of the second type occur in a much less frequent way than events of the first type (it is indeed the case when the size of the sample goes big). The total number of genes carried by a single individual is then closely related to the so-called total external length, which is the sum of all external branch lengths of the tree.

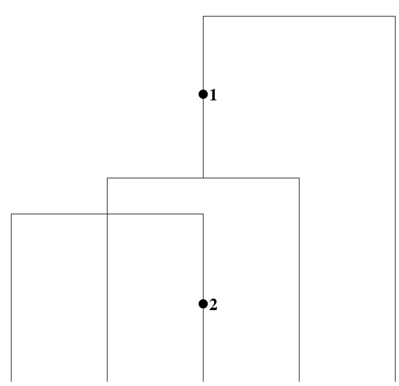

Figure 1. In this genealogical tree, two mutations appear. Mutation 1 is in an internal branch and it is shared by 4 individuals. Mutation 2 is in an external branch so it is carried by 1 individual. In this example, an ancestral gene is also carried by 1 individual. This situation is negligible when the size of the sample is large.

Date: February 7, 2013.

2010 *Mathematics Subject Classification.* 60J70, 60J80, 60J25, 60F05, 92D25.

Key words and phrases. Coalescent process, Bolthausen-Szmitman coalescent, external branch, block counting process, recursive construction, Iksanov-Möhle coupling.

The Bolthausen-Sznitman coalescent (see for instance [6]) is a well-known example of exchangeable coalescents with multiple collisions (see [16, 17] for a proper definition of this type of coalescents). It was first introduced in physics, in order to study spin glasses but it has also been thought as a limiting genealogical model for evolving populations with selective killing at each generation, see for instance [4, 5]. Recently, Berestycki et al. in [3] noted that this coalescent represents the genealogies of the branching brownian motion with absorption.

The Bolthausen-Sznitman coalescent $(\Pi_t, t \geq 0)$, is a continuous time Markov chain with values in the set of partitions of $\mathbb{N}$, starting with an infinite number of blocks/individuals. In order to give a formal description of the Bolthausen-Sznitman coalescent, it is sufficient to give its jump rates. Let $n \in \mathbb{N}$, then the restriction $(\Pi_t^{(n)}, t \geq 0)$ of $(\Pi_t, t \geq 0)$ to $[n] = \{1, \dots, n\}$ is a Markov chain with values in $\mathcal{P}_n$, the set of partitions of $[n]$, with the following dynamics: whenever $\Pi_t^{(n)}$ is a partition consisting of b blocks, any particular k of them merge into one block at rate

$$\lambda_{b,k} = \frac{(k-2)!(b-k)!}{(b-1)!},$$

so the next coalescence event occurs at rate

$$\lambda_b = \sum_{k=2}^{b} \binom{b}{k} \lambda_{b,k} = b - 1. \tag{1}$$

Note that mergers of several blocks into a single block is possible, but multiple mergers do not occur simultaneously. Moreover, this coalescent process is exchangeable, i.e. its law does not change under the effect of a random permutation of labels of its blocks.

One of our aims is to study the total external length of the Bolthausen-Sznitman coalescent. More precisely, we determine the asymptotic behaviour of the total external length $E^{(n)}$ of the Bolthausen-Sznitman coalescent restricted to $\mathcal{P}_n$, when n goes to infinity, and relate it to its total length $L^{(n)}$ (the sum of lengths of all external and internal branches). A first orientation can be gained from coalescents without proper frequencies. For this class Möhle [15] proved that after a suitable scaling the asymptotic distributions of $E^{(n)}$ and $L^{(n)}$ are equal. Now the Bolthausen-Sznitman coalescent does not belong to this class, but it is (loosely speaking) located at the borderline. Also it is known for the Bolthausen-Sznitman coalescent [7] that

$$\frac{(\log n)^2}{n} L^{(n)} - \log n - \log\log n \xrightarrow[n\to\infty]{d} Z, \tag{2}$$

where $\xrightarrow{d}$ denotes convergence in distribution and Z is a strictly stable r.v. with index 1, i.e. its characteristic exponent satisfies

$$\Psi(\theta) = -\log \mathbb{E}\Big[e^{i\theta Z}\Big] = \frac{\pi}{2}|\theta| - i\theta \log|\theta|, \qquad \theta \in \mathbb{R}.$$

In their recent work, Dhersin and Möhle [9] showed that the ratio $E^{(n)}/L^{(n)}$ converges to 1 in probability. Thus one might guess that $E^{(n)}$ satisfies the same asymptotic relation with the same scaling. It is a main result of this paper that such a conjecture is almost, but not completely true.

Let us consider $(\Pi_t^{(n)}, t \geq 0)$. We denote by $U_k^{(n)}$ the size of the k-th jump, i.e the number of blocks that the Markov chain loses in k-th coalescence event. We also denote by $X_k^{(n)}$ for the number of blocks after k coalescence events. Observe that $X_0^{(n)} = n$ and $X_k^{(n)} = X_{k-1}^{(n)} - U_k^{(n)} = n - \sum_{i=1}^k U_i^{(n)}$. Since the merging blocks coalesce into 1, there are

$U_k^{(n)} + 1$ blocks involved in k-th coalescence event and, for $l < X_{k-1}^{(n)}$,

$$\mathbb{P}\Big(U_k^{(n)} = l \Big| X_{k-1}^{(n)} = b\Big) = \frac{\binom{b}{l+1}\lambda_{b,l+1}}{\lambda_b} = \frac{b}{b-1}\frac{1}{l(l+1)}.$$

Let $\tau^{(n)}$ be the number of coalescence events. More precisely

$$\tau^{(n)} = \inf\Big\{k, X_k^{(n)} = 1\Big\}.$$

According to Iksanov and Möhle [12] (see also [8]), $\tau^{(n)}$ satisfies the following asymptotic behaviour

$$\frac{(\log n)^2}{n}\tau^{(n)} - \log n - \log\log n \xrightarrow[n\to\infty]{d} Z. \tag{3}$$

The main result of this paper describes the behaviour of the total internal length $I^{(n)}$, when n goes to ∞. In order to do so, we introduce the r.v. $Y_k^{(n)}$ that represents the number of internal branches after k coalescence events. In other words, it is the number of remaining blocks which have already participated in a coalescence event. Note that at time 0 all branches are external i.e. $Y_0^{(n)} = 0$. Let $(\mathbf{e}_k, k \geq 1)$ be a sequence of i.i.d. standard exponential r.v., also independent from the $X_k^{(n)}$ and $Y_k^{(n)}$, thus from (1), we have

$$I^{(n)} = \sum_{k=1}^{\tau^{(n)}-1} Y_k^{(n)} \frac{\mathbf{e}_k}{X_k^{(n)} - 1}.$$

Our main result is the following weak law of large numbers for $I^{(n)}$. Here $\overset{\mathbb{P}}{\to}$ denotes convergence in probability.

Theorem 1.1. *For the total internal length of the Bolthausen-Sznitman coalescent, we have*

$$\frac{(\log n)^2}{n} I^{(n)} \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

Now noting that $L^{(n)} = I^{(n)} + E^{(n)}$ and using (2) and our main result, we deduce the asymptotic distribution of the total external length $E^{(n)}$.

Corollary 1.2. *For the total external length of the Bolthausen-Sznitman coalescent, we have*

$$\frac{(\log n)^2}{n} E^{(n)} - \log n - \log\log n \xrightarrow[n\to\infty]{d} Z - 1.$$

Observe that the Bolthausen-Sznitman coalescent can be seen as a special case ($\alpha = 1$) of the so-called $Beta(2-\alpha, \alpha)$-coalescent which class is defined for $0 < \alpha < 2$. Möhle's work [15] shows that in the case $0 < \alpha < 1$ the variable $E^{(n)}/n$ converges in law to a random variable defined in terms of a driftless subordinator depending on α. For $1 < \alpha < 2$, we refer to [14] where it is proven that $(E^{(n)} - cn^{2-\alpha})/n^{1/\alpha+1-\alpha}$ converges weakly to a stable r.v. of index α, c being a constant also depending on α (see also [2, 1, 10]). In Kingman's case ($\alpha \to 2$) a logarithmic correction appears and the limit law is normal (see [13]).

The remainder of the paper is structured as follows. In Section 2, we prove our main results using a coupling method which was introduced in [12] that provides more information of the chain $X^{(n)} = (X_k^{(n)}, k \geq 0)$. Finally, Section 3 is devoted to the asymptotic behaviour of the number of mutations appearing in external and internal branches of the Bolthausen-Sznitman coalescent.

2. Proofs

2.1. **A coupling.** In this section, we use the coupling method introduced in [12] in order to study the number of jumps $\tau^{(n)}$.

Let $(V_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with distribution

$$\mathbb{P}(V_1 = k) = \frac{1}{k(k+1)}, \qquad k \geq 1. \tag{4}$$

Note that $\mathbb{P}(V_1 \geq k) = 1/k$. Let $S_n = V_1 + \cdots + V_n$. It is well-known, see for instance [11], that

$$\frac{S_n - n\log n}{n} \xrightarrow[n\to\infty]{d} Z, \tag{5}$$

where Z is the stable random variable that appears in (2). We have the following functional limit result, with a limit, which is certainly a Lévy process.

Lemma 2.1. *The process $(L_n(t), 0 \leq t \leq 1)$ defined by*

$$L_n(t) = \frac{S_{\lfloor nt\rfloor} - \lfloor nt\rfloor \log n}{n}$$

converges weakly in the Skorohod space $\mathcal{D}[0,1]$.

Proof. We first verify that the convergence of finite-dimensional distributions holds. Let $t \geq 0$, from (5), we deduce

$$L_n(t) = \frac{S_{\lfloor nt\rfloor} - \lfloor nt\rfloor \log(nt) + \lfloor nt\rfloor \log t}{n} \xrightarrow[n\to\infty]{d} tZ + t\log t.$$

Similarly, if we take $s \leq t$, then

$$\Big(L_n(s), L_n(t) - L_n(s)\Big) \xrightarrow[n\to\infty]{d} (Z_1, Z_2),$$

where Z_1 and Z_2 are independent random variables distributed as $sZ + s\log s$ and $(t-s)Z + (t-s)\log(t-s)$, respectively. The mapping theorem implies that

$$\Big(L_n(s), L_n(t)\Big) \xrightarrow[n\to\infty]{d} (Z_1, Z_1 + Z_2).$$

A set of three or more time points can be treated in the same way, and hence the finite-dimensional distributions converge properly.

We now check tightness via Aldous' criterion. Let T_n be a L_n-stopping time and (θ_n) a sequence of positive numbers such that $\theta_n \to 0$, as n increases. Then for $\varepsilon > 0$, we have

$$\begin{aligned}\mathbb{P}\Big(\Big|L_n(T_n+\theta_n) - L_n(T_n)\Big| = \varepsilon\Big) &\leq \mathbb{P}\Big(\Big|L_n(\theta_n)\Big| \geq \varepsilon\Big)\\ &= \mathbb{P}\left(\left|\frac{S_{\lfloor n\theta_n\rfloor} - \lfloor n\theta_n\rfloor \log(n\theta_n)}{n\theta_n}\theta_n + \frac{\lfloor n\theta_n\rfloor \log\theta_n}{n}\right| \geq \varepsilon\right)\end{aligned}$$

which converges to 0. This completes the proof. □

In what follows, we use the following notation. For a stochastic process $(Z_n, n \geq 0)$ and a function $c(n)$ write $Z_n = O_p(c(n))$ as $n \to \infty$, if $Z_n/c(n)$ is stochastically bounded as $n \to \infty$, i.e. if $\lim_{x\to\infty} \limsup_{n\to\infty} \mathbb{P}(|Z_n| > xc(n)) = 0$. We also write $Z_n = o_p(c(n))$ as $n \to \infty$, if $Z_n/c(n)$ goes to 0 in probability.

From the above result, we deduce

$$\sup_{1\le k\le n}\Big|S_k - k\log n\Big| = O_p(n). \tag{6}$$

We now define recursively $(\rho(k))_{k\ge 0}$, a sequence of stopping times such that $\rho(0)=0$ and

$$\rho(k+1) = \inf\left\{ i > \rho(k), V_i + \sum_{j=1}^{k} V_{\rho(j)} < n \right\}$$

with the convention $\inf\{\emptyset\} = \infty$. In other words, the sequence $(\rho(k))_{k\ge 1}$ is the collection of indices of the r.v.'s V_i such that their sum does not exceed $n-1$. It is proved in [12] that $\tau^{(n)}$ and $\sup\{k, \rho(k) < \infty\}$ are equal in law, and that the terms of the block-counting Markov chain of the Bolthausen-Sznitman coalescent can be represented as $X_0^{(n)} = n$, and

$$X_k^{(n)} = n - \sum_{i=1}^{k} V_{\rho(i)}.$$

Next, we define

$$\sigma^{(n)} = \inf\{k, \rho(k) > k\},$$

the first time that the random walk meets or exceeds n, and

$$\theta_\gamma^{(n)} = \tau^{(n)} - \frac{n}{(\log n)^{1+\gamma}}, \qquad \gamma \in (0,\infty], \tag{7}$$

with the convention $\theta_\infty^{(n)} = \tau^{(n)}$. Our first result allows us to consider the random walk instead of the process of disappearing blocks until time $\theta_\gamma^{(n)}$.

Proposition 2.2. *Let $0<\gamma<\gamma'\le\infty$. Then as n goes to ∞, we have*

$$\mathbb{P}\Big(\theta_\gamma^{(n)} < \sigma^{(n)}\Big) \to 1 \qquad \text{and} \qquad \frac{(\log n)^\gamma}{n} X^{(n)}_{\theta_\gamma^{(n)}} \xrightarrow{\mathbb{P}} 1. \tag{8}$$

Moreover,

$$\sup_{1\le k\le \theta_\gamma^{(n)}} \left| \frac{X_k^{(n)}}{\theta_{\gamma'}^{(n)} - k} - \log n \right| = o_p(\log n), \tag{9}$$

for n sufficiently large.

In order to prove this proposition, we first show that a similar result holds for the family of stopping times

$$\eta_{c,\gamma}^{(n)} = \inf\left\{k, X_k^{(n)} < \frac{cn}{(\log n)^\gamma}\right\},$$

where c is a positive constant, and then note that for $\epsilon>0$,

$$\mathbb{P}\Big(\eta_{1-\epsilon,\gamma}^{(n)} \le \theta_\gamma^{(n)} \le \eta_{1+\epsilon,\gamma}^{(n)}\Big) \xrightarrow[n\to\infty]{} 1.$$

Hence, the proof of Proposition 2.2 relies on the following Lemma.

Lemma 2.3. *Let $0<\gamma<\infty$. Then as n goes to ∞, we have*

$$\mathbb{P}\Big(\eta_{c,\gamma}^{(n)} < \sigma^{(n)}\Big) \to 1 \qquad \text{and} \qquad \frac{(\log n)^\gamma}{cn} X^{(n)}_{\eta_{c,\gamma}^{(n)}} \xrightarrow{\mathbb{P}} 1.$$

Moreover,

$$\sup_{1\le k\le \eta_{c,\gamma}^{(n)}}\left|\frac{X_k^{(n)}}{\tau^{(n)}-k}-\log n\right| = o_p(\log n),$$

for n sufficiently large.

Proof. First we prove the result for $\gamma<1$. Observe, from (4), that for any $\varepsilon>0$

$$\mathbb{P}\left(V_k<\frac{n}{(\log n)^{1-\varepsilon}},\ \text{for all } k\le 2n/\log n\right)\ge\left(1-\frac{(\log n)^{1-\varepsilon}}{n}\right)^{2n/\log n}\xrightarrow[n\to\infty]{}1.$$

Now, since $\mathbb{P}(\tau^{(n)}\le\frac{2n}{\log n})\to 1$, as n increases (which follows from (3)), we get

$$\sup_{1\le k\le\tau^{(n)}} V_k = O_p\left(\frac{n}{(\log n)^{1-\varepsilon}}\right). \tag{10}$$

For simplicity, we write $\eta^{(n)}$ instead of $\eta_{c,\gamma}^{(n)}$. Then, it follows

$$\begin{aligned}\mathbb{P}\Big(\sigma^{(n)}\le\eta^{(n)}\Big) &= \mathbb{P}\Big(V_k\ge X_{k-1}^{(n)},\quad\text{for some } k\le\eta^{(n)}\Big)\\ &\le \mathbb{P}\left(V_k\ge\frac{cn}{(\log n)^\gamma},\quad\text{for some } k\le\tau^{(n)}\right)\\ &= \mathbb{P}\left(\sup_{1\le k\le\tau^{(n)}}V_k\ge\frac{cn}{(\log n)^\gamma}\right),\end{aligned}$$

thus if we take $\varepsilon\in(0,1-\gamma)$ in (10), we deduce

$$\mathbb{P}\Big(\sigma^{(n)}\le\eta^{(n)}\Big)\xrightarrow[n\to\infty]{}0. \tag{11}$$

On the event $\{\sigma^{(n)}>\eta^{(n)}\}$, it is clear that

$$\sup_{1\le k\le\eta^{(n)}}\left|\frac{X_k^{(n)}}{X_{k-1}^{(n)}}-1\right| = \sup_{1\le k\le\eta^{(n)}}\frac{V_k}{X_{k-1}^{(n)}}\le\frac{(\log n)^\gamma}{cn}\sup_{1\le k\le\tau^{(n)}}V_k.$$

Hence, from (10) and (11), we obtain

$$\sup_{1\le k\le\eta^{(n)}}\left|\frac{X_k^{(n)}}{X_{k-1}^{(n)}}-1\right| = o_p(1). \tag{12}$$

In particular, since $X_{\eta^{(n)}}^{(n)}\le\frac{cn}{(\log n)^\gamma}\le X_{\eta^{(n)}-1}^{(n)}$, we get

$$\frac{(\log n)^\gamma}{cn}X_{\eta^{(n)}}^{(n)}\xrightarrow[n\to\infty]{\mathbb{P}}1. \tag{13}$$

Next, we note

$$X_k^{(n)}-(\tau^{(n)}-k)\log n = X_k^{(n)}-n+k\log\left(\frac{2n}{\log n}\right)+(n-\tau^{(n)}\log n)+k\log\left(\frac{\log n}{2}\right),$$

and from (3), it is clear

$$\frac{(\log n)^2}{n}\tau^{(n)} = \log n + O_p(\log\log n).$$

Then on the event $\{\eta^{(n)} < \sigma^{(n)}, \eta^{(n)} < \frac{2n}{\log n}\}$, it follows from (6) that

$$\sup_{1\le k\le \eta^{(n)}} \left| X_k^{(n)} - (\tau^{(n)} - k)\log n \right| \le \sup_{1\le k\le 2n/\log n} \left| S_k - k\log\left(\frac{2n}{\log n}\right)\right| + O_p\left(\frac{n\log\log n}{\log n}\right)$$
$$= O_p\left(\frac{n\log\log n}{\log n}\right).$$

Finally using (13) and the strong Markov property for $\tilde{X}_k^{(n)} = X_{k+\eta^{(n)}}^{(n)}$, we deduce

$$\tau^{(n)} - \eta^{(n)} = \tau^{(\tilde{X}_0^{(n)})} = \frac{X_{\eta^{(n)}}^{(n)}}{\log X_{\eta^{(n)}}^{(n)}}(1+o_p(1)) = \frac{cn}{(\log n)^{1+\gamma}}(1+o_p(1)). \tag{14}$$

Then, putting all the pieces together, we get

$$\sup_{1\le k\le \eta^{(n)}} \left| \frac{X_k^{(n)}}{\tau^{(n)} - k} - \log n \right| \le \frac{\sup_{1\le k\le \eta^{(n)}} \left| X_k^{(n)} - (\tau^{(n)} - k)\log n\right|}{\tau^{(n)} - \eta^{(n)}} = O_p((\log n)^\gamma \log\log n),$$

and since $\gamma < 1$,

$$\sup_{1\le k\le \eta^{(n)}} \left| \frac{X_k^{(n)}}{\tau^{(n)} - k} - \log n \right| = o_p(\log n). \tag{15}$$

Next, we will prove (11), (12), (13), (14) and (15) for any $\gamma > 0$. We show that this claim holds for $\gamma \le p/2$ for any $p \in \mathbb{N}$, using induction on p. The proof for $p = 1$ has just been done.

For the induction step suppose that the asymptotics in (11) to (15) hold for $\gamma \le p/2$. For simplicity, we write $\hat{\eta}^{(n)} = \eta_{c,p/2}^{(n)}$. The idea is to use the strong Markov property at the stopping time $\hat{\eta}^{(n)}$ and apply the above results for $\gamma < 1$ to the Markov chain $\hat{X}_k^{(n)} = X_{k+\hat{\eta}^{(n)}}^{(n)}$ started at $\hat{n} = X_{\hat{\eta}^{(n)}}^{(n)}$ (instead of $n = X_0^{(n)}$). Define the family of stopping times

$$\zeta^{(n)} = \inf\left\{k, X_k^{(n)} < \frac{\hat{n}}{(\log \hat{n})^{2/3}}\right\}.$$

Observe that $\zeta^{(n)} = \hat{\eta}^{(n)} + \eta_{1,\frac{2}{3}}^{(\hat{n})}$. Hence, using the strong Markov property at the stopping time $\hat{\eta}^{(n)}$ and the behaviour in (13), with $\gamma = 2/3$, we get

$$\frac{(\log X_{\hat{\eta}^{(n)}}^{(n)})^{\frac{2}{3}}}{X_{\hat{\eta}^{(n)}}^{(n)}} X_{\zeta^{(n)}}^{(n)} \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

Then, from this asymptotic behaviour and the induction hypothesis taken in (13),

$$\frac{(\log n)^{\frac{2}{3}+\frac{p}{2}}}{cn} X_{\zeta^{(n)}}^{(n)} \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

From this behaviour, from (13) and from

$$\frac{(\log n)^\gamma}{cn} X_{\eta^{(n)}}^{(n)} < 1 \le \frac{(\log n)^\gamma}{cn} X_{\eta^{(n)}-1}^{(n)}$$

we obtain for $p/2 < \gamma \le (p+1)/2$

$$\mathbb{P}\Big(\hat{\eta}^{(n)} < \eta^{(n)} \le \zeta^{(n)}\Big) \xrightarrow[n\to\infty]{} 1. \tag{16}$$

Now, on the event $\{\sigma^{(n)} > \hat{\eta}^{(n)}\}$, using the strong Markov property at $\hat{\eta}^{(n)}$ and (11) with the initial state $X^{(n)}_{\hat{\eta}^{(n)}}$, we get

$$\mathbb{P}\Big(\sigma^{(n)} \le \zeta^{(n)} \big| \sigma^{(n)} > \hat{\eta}^{(n)}\Big) \xrightarrow[n\to\infty]{} 0.$$

The induction hypothesis gives $\mathbb{P}(\sigma^{(n)} > \hat{\eta}^{(n)}) \to 1$, as n goes to ∞. These two facts together lead to

$$\mathbb{P}\Big(\sigma^{(n)} > \eta^{(n)}\Big) \xrightarrow[n\to\infty]{} 1,$$

for $\gamma \in (p/2, (p+1)/2]$.

From (12) and again the strong Markov property at $\hat{\eta}^{(n)}$, we get

$$\sup_{\hat{\eta}^{(n)} \le k \le \zeta^{(n)}} \left| \frac{X^{(n)}_k}{X^{(n)}_{k-1}} - 1 \right| = o_p(1).$$

From the above behaviour, (16) and the induction hypothesis, we have

$$\sup_{1 \le k \le \eta^{(n)}} \left| \frac{X^{(n)}_k}{X^{(n)}_{k-1}} - 1 \right| = o_p(1),$$

for $\gamma \in (p/2, (p+1)/2]$. Again the strong Markov property together with the above behaviour give us for all $\gamma \in (p/2, (p+1)/2]$,

$$\tau^{(n)} - \eta^{(n)} = \frac{cn}{(\log n)^{1+\gamma}}(1 + o_p(1)).$$

From (15) and the strong Markov property, we get

$$\sup_{\hat{\eta}^{(n)} \le k \le \zeta^{(n)}} \left| \frac{X^{(n)}_k}{\tau^{(n)} - k} - \log X^{(n)}_{\hat{\eta}^{(n)}} \right| = o_p\Big(\log X^{(n)}_{\hat{\eta}^{(n)}}\Big).$$

We know from the induction hypothesis that $\log X^{(n)}_{\hat{\eta}^{(n)}} \sim \log n$, as n goes to ∞. We then obtain, from (16) and using again the induction hypothesis,

$$\sup_{1 \le k \le \eta^{(n)}} \left| \frac{X^{(n)}_k}{\tau^{(n)} - k} - \log n \right| = o_p(\log n),$$

for all $\gamma \in (p/2, (p+1)/2]$. Hence the induction is complete and the behaviour in (11) to (15) hold for any $\gamma > 0$. $\square$

Proof of Proposition 2.2. We first recall the definition of $\theta^{(n)}_\gamma$ in (7) and define $\eta^{(n)}_- = \eta^{(n)}_{1-\varepsilon,\gamma}$ and $\eta^{(n)}_+ = \eta^{(n)}_{1+\varepsilon,\gamma}$. From (14), it is clear

$$\mathbb{P}\Big(\eta^{(n)}_+ \le \theta^{(n)}_\gamma \le \eta^{(n)}_-\Big) \xrightarrow[n\to\infty]{} 1,$$

and from (11), we deduce

$$\mathbb{P}\Big(\sigma^{(n)} > \eta^{(n)}_-\Big) \xrightarrow[n\to\infty]{} 1.$$

Thus the first asymptotic behaviour in (8) holds. Also note

$$\mathbb{P}\Big(X^{(n)}_{\eta^{(n)}_-} \le X^{(n)}_{\theta^{(n)}_\gamma} \le X^{(n)}_{\eta^{(n)}_+}\Big) \xrightarrow[n\to\infty]{} 1,$$

then the second asymptotic behaviour in (8) follows from (13).

From (15), we get

$$\sup_{1\le k\le \theta_\gamma^{(n)}} \left| \frac{X_k^{(n)}}{\tau^{(n)}-k} - \log n \right| = o_p(\log n)$$

which gives (9) for $\gamma' = \infty$. Also

$$\sup_{1\le k\le \theta_\gamma^{(n)}} \left| \frac{\theta_{\gamma'}^{(n)}-k}{\tau^{(n)}-k} - 1 \right| = \sup_{1\le k\le \theta_\gamma^{(n)}} \left| \frac{\tau^{(n)}-\theta_{\gamma'}^{(n)}}{\tau^{(n)}-k} \right| \le \frac{\tau^{(n)}-\theta_{\gamma'}^{(n)}}{\tau^{(n)}-\theta_{\gamma}^{(n)}} = \frac{(\log n)^\gamma}{(\log n)^{\gamma'}}(1+o_p(1)).$$

This give us (9). This completes the proof. □

2.2. **Proof of Theorem 1.1.** We first define

$$\tilde{I}^{(n)} = \sum_{k=1}^{\tau^{(n)}-1} \frac{Y_k^{(n)}}{X_k^{(n)}},$$

which is obtained by replacing the exponential random variables $\mathbf{e}_k$'s by their mean and approximating the denominator. Similarly, we define

$$\hat{I}^{(n)} = \sum_{k=1}^{\tau^{(n)}-1} \frac{\mathbb{E}[Y_k^{(n)}|X^{(n)}]}{X_k^{(n)}},$$

which is obtained by replacing the random variables $Y_k^{(n)}$ by its conditional expectation. This new formulation is of interest. Indeed, similar as in [14] it is possible to determine $\hat{I}^{(n)}$ via a recursive formula. Let $Z_k^{(n)}$ be the number of external branches after k jumps, $k \ge 1$, and we take conditional expectation to each $Z_k^{(n)}$ with respect to $X^{(n)}$ and $Z_{k-1}^{(n)}$. Observe that $Z_{k-1}^{(n)} - Z_k^{(n)}$ is the number of external branches which participate to k-th coalescent event. Hence, this random variable is distributed as an hypergeometric r.v. with parameters $X_{k-1}^{(n)}$, $Z_{k-1}^{(n)}$ and $1+U_k^{(n)}$. It is then clear

$$\mathbb{E}\left[Z_k^{(n)} \middle| X^{(n)}, Z_{k-1}^{(n)}\right] = Z_{k-1}^{(n)} - \left(1+U_k^{(n)}\right)\frac{Z_{k-1}^{(n)}}{X_{k-1}^{(n)}},$$

then

$$\mathbb{E}\left[Z_k^{(n)} \middle| X^{(n)}\right] = \mathbb{E}\left[Z_{k-1}^{(n)} \middle| X^{(n)}\right]\frac{X_k^{(n)}-1}{X_{k-1}^{(n)}},$$

and

$$\frac{\mathbb{E}\left[Z_k^{(n)} \middle| X^{(n)}\right]}{X_k^{(n)}} = \prod_{i=1}^{k}\left(1-\frac{1}{X_i^{(n)}}\right).$$

Finally, since $Y_k^{(n)} = X_k^{(n)} - Z_k^{(n)}$, it follows

$$\hat{I}^{(n)} = \sum_{k=1}^{\tau^{(n)}-1}\left(1-\prod_{i=1}^{k}\left(1-\frac{1}{X_i^{(n)}}\right)\right). \tag{17}$$

This last expression is a good way to understand the asymptotic behaviour of the total internal branch.

The following lemma provides the asymptotic behaviour of $\hat{I}^{(n)}$.

Lemma 2.4. *As n goes to ∞,*

$$\frac{(\log n)^2}{n}\hat{I}^{(n)} \to 1,$$

in probability.

Proof. Let $\varepsilon > 0$ and take $\theta_\gamma^{(n)}$ as in (7). We also let $\theta_-^{(n)} = \lfloor \theta_{1-\varepsilon}^{(n)} \rfloor$ and $\theta_+^{(n)} = \lfloor \theta_{1+\varepsilon}^{(n)} \rfloor$ and consider $\hat{I}^{(n)}$ as it is given in (17). We now split $\hat{I}^{(n)}$ in two parts, as follows

$$\hat{I}_1^{(n)} = \sum_{k=1}^{\theta_+^{(n)}-1} \left(1 - \prod_{i=1}^{k}\left(1 - \frac{1}{X_i^{(n)}}\right)\right),$$

and

$$\hat{I}_2^{(n)} = \sum_{k=\theta_+^{(n)}}^{\tau^{(n)}-1} \left(1 - \prod_{i=1}^{k}\left(1 - \frac{1}{X_i^{(n)}}\right)\right).$$

Note that

$$\hat{I}_2^{(n)} \le \tau^{(n)} - \theta_+^{(n)} \le \frac{n}{(\log n)^{2+\varepsilon}} + 1,$$

which implies that

$$\frac{(\log n)^2}{n}\hat{I}_2^{(n)} \xrightarrow[n\to\infty]{} 0, \qquad \text{almost surely.}$$

Then it is enough to study the behaviour of $\hat{I}_1^{(n)}$. In order to do so, we first note

$$\sum_{i=1}^{k}\frac{1}{X_i^{(n)}} - \sum_{j=2}^{k}\sum_{i=1}^{j-1}\frac{1}{X_i^{(n)}X_j^{(n)}} \le 1 - \prod_{i=1}^{k}\left(1 - \frac{1}{X_i^{(n)}}\right) \le \sum_{i=1}^{k}\frac{1}{X_i^{(n)}}.$$

(This can be viewed as two Bonferroni inequalities for independent events with entrance probabilities $1/X_i^{(n)}$.)

On the one hand,

$$\sum_{k=1}^{\theta_+^{(n)}-1}\sum_{i=1}^{k}\frac{1}{X_i^{(n)}} = \sum_{i=1}^{\theta_+^{(n)}-1}\frac{\theta_+^{(n)} - i}{X_i^{(n)}}$$

and thus

$$\sum_{i=1}^{\theta_-^{(n)}-1}\frac{\theta_+^{(n)} - i}{X_i^{(n)}} \le \sum_{k=1}^{\theta_+^{(n)}-1}\sum_{i=1}^{k}\frac{1}{X_i^{(n)}} \le \sum_{i=1}^{\theta_+^{(n)}-1}\frac{\tau^{(n)} - i}{X_i^{(n)}}.$$

From (9), we get

$$\frac{1}{\log n}(\theta_-^{(n)} - 1)(1 + o_p(1)) \le \sum_{k=1}^{\theta_+^{(n)}-1}\sum_{i=1}^{k}\frac{1}{X_i^{(n)}} \le \frac{1}{\log n}(\theta_+^{(n)} - 1)(1 + o_p(1)).$$

From the fact that $\theta_-^{(n)}, \theta_+^{(n)} \sim \tau^{(n)} \sim n/\log n$, as $n \to \infty$, we deduce

$$\sum_{k=1}^{\theta_+^{(n)}-1}\sum_{i=1}^{k}\frac{1}{X_i^{(n)}} = \frac{n}{(\log n)^2}(1 + o_p(1)).$$

On the other hand by inverting the sums, we obtain

$$\sum_{k=1}^{\theta_+^{(n)}-1}\sum_{j=2}^{k}\sum_{i=1}^{j-1}\frac{1}{X_i^{(n)}X_j^{(n)}}=\sum_{j=2}^{\theta_+^{(n)}-1}\sum_{k=j}^{\theta_+^{(n)}-1}\sum_{i=1}^{j-1}\frac{1}{X_i^{(n)}X_j^{(n)}}=\sum_{j=2}^{\theta_+^{(n)}-1}\frac{\theta_+^{(n)}-j}{X_j^{(n)}}\sum_{i=1}^{j-1}\frac{1}{X_i^{(n)}}.$$

Using (9), we obtain

$$\sum_{k=1}^{\theta_+^{(n)}-1}\sum_{j=2}^{k}\sum_{i=1}^{j-1}\frac{1}{X_i^{(n)}X_j^{(n)}}\leq\sum_{j=2}^{\theta_+^{(n)}-1}\frac{\tau^{(n)}-j}{X_j^{(n)}}\sum_{i=1}^{j-1}\frac{1}{X_i^{(n)}}\leq\frac{1+o_p(1)}{\log n}\sum_{j=1}^{\theta_+^{(n)}-1}\sum_{i=1}^{j}\frac{1}{X_i^{(n)}}.$$

and finally

$$\sum_{k=1}^{\theta_+^{(n)}-1}\sum_{j=2}^{k}\sum_{i=1}^{j-1}\frac{1}{X_i^{(n)}X_j^{(n)}}\leq\frac{n}{(\log n)^3}(1+o_p(1)).$$

Putting all the pieces together give us

$$\frac{(\log n)^2}{n}\hat{I}_1^{(n)}\xrightarrow[n\to\infty]{\mathbb{P}}1,$$

which ends the proof. □

In order to prove Theorem 1.1, we just need to control our approximation. This is the aim of the next two lemmas.

Lemma 2.5. *As n goes to ∞,*

$$I^{(n)}-\tilde{I}^{(n)}=O_P(\sqrt{n}).$$

Proof. Recall that $X^{(n)}$ denotes the Markov chain $(X_k^{(n)},k\geq 0)$. A simple computation gives us

$$\left(I^{(n)}-\tilde{I}^{(n)}\right)=\sum_{k=1}^{\tau^{(n)}-1}Y_k^{(n)}\frac{\mathbf{e}_k-1}{X_k^{(n)}}+\sum_{k=1}^{\tau^{(n)}-1}Y_k^{(n)}\frac{\mathbf{e}_k}{X_k^{(n)}(X_k^{(n)}-1)}.$$

Conditionally on $X^{(n)},Y^{(n)}$, the random variables $Y_k^{(n)}\frac{\mathbf{e}_k-1}{X_k^{(n)}}$ are independent with zero mean. This implies

$$\mathbb{E}\left[\left(\sum_{k=1}^{\tau^{(n)}-1}Y_k^{(n)}\frac{\mathbf{e}_k-1}{X_k^{(n)}}\right)^2\middle|X^{(n)},Y^{(n)}\right]=\sum_{k=1}^{\tau^{(n)}-1}\left(\frac{Y_k^{(n)}}{X_k^{(n)}}\right)^2\leq\tau^{(n)}\leq n,$$

where the inequality follows from the fact that $Y_k^{(n)}\leq X_k^{(n)}$ a.s. Chebychev's inequality implies

$$\sum_{k=1}^{\tau^{(n)}-1}Y_k^{(n)}\frac{\mathbf{e}_k-1}{X_k^{(n)}}=O_P(\sqrt{n}).$$

Again using that $Y_k^{(n)}\leq X_k^{(n)}$ a.s., we get

$$\sum_{k=0}^{\tau^{(n)}-1}Y_k^{(n)}\frac{\mathbf{e}_k}{X_k^{(n)}(X_k^{(n)}-1)}\leq\sum_{k=0}^{\tau^{(n)}-1}\frac{\mathbf{e}_k}{(X_k^{(n)}-1)}\leq\sum_{k=0}^{n-1}\frac{\mathbf{e}_k}{(k-1)}.$$

It is a classical result (used also for the total length of Kingman coalescent) that

$$\frac{1}{\log n}\sum_{k=0}^{n-1}\frac{\mathbf{e}_k}{(k-1)}\xrightarrow[n\to\infty]{\mathbb{P}}1,$$

which implies that

$$\sum_{k=0}^{\tau^{(n)}-1}Y_k^{(n)}\frac{\mathbf{e}_k}{X_k^{(n)}(X_k^{(n)}-1)}=O_P(\log n).$$

This completes the proof. $\square$

Lemma 2.6. *As n goes to ∞*

$$\tilde{I}^{(n)}-\hat{I}^{(n)}=O_P(\sqrt{n}).$$

Proof. We proceed similar as in [14]. Recall that $Z_k^{(n)}$ is the number of external branches after k coalescing events. Since $Y_k^{(n)}=X_k^{(n)}-Z_k^{(n)}$,

$$\tilde{I}^{(n)}-\hat{I}^{(n)}=-\sum_{k=1}^{\tau^{(n)}-1}\frac{Z_k^{(n)}-\mathbb{E}[Z_k^{(n)}|X^{(n)}]}{X_k^{(n)}}$$

Also recall that $Z_k^{(n)}-Z_{k-1}^{(n)}$ has a conditional hypergeometric distribution, given $X^{(n)},Z_{k-1}^{(n)}$. Therefore

$$Z_k^{(n)}=Z_{k-1}^{(n)}-(U_k^{(n)}+1)\frac{Z_{k-1}^{(n)}}{X_{k-1}^{(n)}}-H_k^{(n)}=Z_{k-1}^{(n)}\frac{X_k^{(n)}-1}{X_{k-1}^{(n)}}-H_k^{(n)},$$

where $H_k^{(n)}$ denotes a random variable with conditional hypergeometric distribution with parameters $X_{k-1}^{(n)}$, $Z_{k-1}^{(n)}$ and $1+U_k^{(n)}$ as above, centered at its (conditional) expectation. For

$$D_k^{(n)}=Z_k^{(n)}-\mathbb{E}[Z_k^{(n)}|X^{(n)}]$$

it follows

$$D_k^{(n)}=D_{k-1}^{(n)}\frac{X_k^{(n)}-1}{X_{k-1}^{(n)}}-H_k^{(n)}\ .$$

Iterating this linear recursion we obtain because of $D_0^{(n)}=0$

$$\frac{D_k^{(n)}}{X_k^{(n)}}=-\sum_{j=1}^{k}\frac{H_j^{(n)}}{X_j^{(n)}}\prod_{i=j+1}^{k}\Big(1-\frac{1}{X_i^{(n)}}\Big)$$

and consequently

$$\tilde{I}^{(n)}-\hat{I}^{(n)}=\sum_{k=1}^{\tau^{(n)}-1}\sum_{j=1}^{k}\frac{H_j^{(n)}}{X_j^{(n)}}\prod_{i=j+1}^{k}\Big(1-\frac{1}{X_i^{(n)}}\Big)=\sum_{j=1}^{\tau^{(n)}-1}\frac{H_j^{(n)}}{X_j^{(n)}}\sum_{k=j}^{\tau^{(n)}-1}\prod_{i=j+1}^{k}\Big(1-\frac{1}{X_i^{(n)}}\Big).$$

Now, since the $H_k^{(n)}$ are centered hypergeometric variables, they are uncorrelated, given $X^{(n)}$. Also from the formula for the variance of a hypergeometric distribution

$$\mathbb{E}[(H_j^{(n)})^2|X^{(n)},Z_{j-1}^{(n)}]\le(U_j^{(n)}+1)\frac{Z_{j-1}^{(n)}}{X_{j-1}^{(n)}}$$

thus, since $Z_{j-1}^{(n)} \le X_{j-1}^{(n)}$ a.s.

$$\mathbb{E}[(H_j^{(n)})^2 | X^{(n)}] \le (U_j^{(n)} + 1).$$

Putting everything together we obtain

$$\mathbb{E}[(\hat{I}^{(n)} - \tilde{I}^{(n)})^2 | X^{(n)}] \le \sum_{j=1}^{\tau^{(n)}-1} \frac{U_j^{(n)} + 1}{(X_j^{(n)})^2} \Big(\sum_{k=j}^{\tau^{(n)}-1} \prod_{i=j+1}^{k} \Big(1 - \frac{1}{X_i^{(n)}}\Big)\Big)^2.$$

The product can be estimated by 1, thus

$$\mathbb{E}[(\hat{I}^{(n)} - \tilde{I}^{(n)})^2 | X^{(n)}] \le \sum_{j=1}^{\tau^{(n)}-1} \frac{U_j^{(n)} + 1}{(X_j^{(n)})^2} (\tau^{(n)} - j)^2.$$

By means of $\tau^{(n)} - j \le X_j^{(n)}$

$$\mathbb{E}[(\hat{I}^{(n)} - \tilde{I}^{(n)})^2 | X^{(n)}] \le \sum_{j=1}^{\tau^{(n)}-1} (U_j^{(n)} + 1) \le n + \tau_n \le 2n.$$

Now an application of Chebychev's inequality gives the claim. □

3. Application to population genetics

Let us now suppose that mutations occur along genealogical trees according to a Poisson process of intensity μ. We write by $M^{(n)}$ for the total number of mutations in the Bolthausen-Sznitman n-coalescent. The Poissonian representation implies that, conditionally on $L^{(n)}$, $M^{(n)}$ is distributed as a Poisson r.v. with parameter $\mu L^{(n)}$. Mutations can be divided as external and internal according to the type of the branches where they appear and we denote them by $M_E^{(n)}$ and $M_I^{(n)}$, respectively.

Proposition 3.1. *As n goes to ∞,*

$$\frac{(\log n)^2}{n} M_I^{(n)} \to \mu,$$

in probability and

$$\frac{(\log n)^2}{n} M_E^{(n)} - \mu \log n - \mu \log\log n \to \mu(Z - 1),$$

in distribution.

Proof. Let $N = (N_t, t \ge 0)$ be a Poisson process with parameter μ. We first note that conditionally on $I^{(n)}$, $M_I^{(n)}$ has the same distribution as $N_{I^{(n)}}$. This implies

$$\mathbb{E}\Big[M_I^{(n)}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[M_I^{(n)} | I^{(n)}\Big]\Big] = \mu \mathbb{E}\Big[I^{(n)}\Big].$$

Since $I^{(n)} \to \infty$ a.s., thanks Theorem 1.1, we deduce that $N_{I^{(n)}}/I^{(n)} \to \mu$ in probability and

$$\frac{M_I^{(n)}}{\mathbb{E}\Big[M_I^{(n)}\Big]} \stackrel{d}{=} \frac{N_{I^{(n)}}}{\mu I^{(n)}} \frac{I^{(n)}}{\mathbb{E}\Big[I^{(n)}\Big]} \xrightarrow[n\to\infty]{\mathbb{P}} 1,$$

thanks to Theorem 1.1. Therefore, the first result follows from $\mathbb{E}[M_I^{(n)}] = \mu \mathbb{E}[I^{(n)}] \sim \mu n/(\log n)^2$, as $n \to \infty$.

To get the second part of this proposition, we just need to observe that $M^{(n)} = M_I^{(n)} + M_E^{(n)}$ satisfies (see Corollary 6.2 of [7])

$$\frac{(\log n)^2}{n} M^{(n)} - \mu \log n - \mu \log\log n \xrightarrow[n\to\infty]{} \mu Z,$$

in distribution. □

Acknowledgement. G.K. acknowledges support by the DFG priority program SPP-1590 'Probabilistic structures in Evolution. J.C.P. acknowledges support by CONACYT, No. 128896.

References

[1] J. Berestycki, N. Berestycki and J. Schweinsberg , Beta-coalescents and continuous stable random trees, *Ann. Probab.* **35** (2007), no. 5, 1835–1887. MR 2349577 (2009d:60244)
[2] J. Berestycki, N. Berestycki and J. Schweinsberg , Small-time behavior of Beta-coalescents, *Ann. Inst. H. Poincaré Probab. Stat.* **44** (2008), no. 2, 214–238. MR MR2446321
[3] J. Berestycki, N. Berestycki and J. Schweinsberg, The genealogy of branching Brownian motion with absorption, To appear in *Ann. Probab.*
[4] É. Brunet, B. Derrida, A.H. Müller and S. Munier, Noisy traveling waves: effect of selection on genealogies, *Europhys. Lett.* **76** (2006), no. 1, 1–7. MR MR2299937 (2007m:82070)
[5] É. Brunet, B. Derrida, A.H. Müller and S. Munier, Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization, *Phys. Rev. E (3)* **76** (2007), no. 4, 041104, 20. MR MR2365627 (2008k:82066)
[6] E. Bolthausen and A.-S. Sznitman, On Ruelle's probability cascades and an abstract cavity method, *Comm. Math. Phys.* **197** (1998), no. 2, 247–276. MR 1652734 (99k:60244)
[7] M. Drmota, A. Iksanov, M. Möhle and U Röler, Asymptotic results concerning the total branch length of the Bolthausen-Sznitman coalescent, *Stochastic Process. Appl.* **117** (2007), no. 10, 1404–1421.
[8] M. Drmota, A. Iksanov, M. Möhle and U Röler, A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree, *Random Structures Algorithms* **34** (2009), no. 3, 319–336. MR 2504401 (2010g:60018)
[9] J.-S. Dhersin and M. Möhle, On the external branches of coalescent processes with multiple collisions with an emphasis on the bolthausen-sznitman coalescent, Preprint available on *http://arxiv.org/abs/1209.3380*
[10] J.-S. Dhersin and L. Yuan, Asympotic behavior of the total length of external branches for beta-coalescents, Preprint available on *http://arxiv.org/abs/1202.5859*
[11] J. L. Geluk and L. de Haan, Stable probability distributions and their domains of attraction: a direct approach, *Probab. Math. Statist.* **20** (2000), no. 1, Acta Univ. Wratislav. No. 2246, 169–188. MR 1785245 (2001g:60031)
[12] A. Iksanov and M. Möhle, A probabilistic proof of a weak limit law for the number of cuts needed to isolate the root of a random recursive tree, *Electron. Comm. Probab.* **12** (2007), 28–35. MR 2407414 (2009e:05076)
[13] S. Janson and G. Kersting, On the total external length of the Kingman coalescent, *Electron. J. Probab.* **16** (2011), 2203–2218.
[14] G. Kersting, I. Stanciu and A. Wakolbinger, The total external branch length of beta- coalescents. Preprint available on *http://arxiv.org/abs/1212.6070*
[15] M. Möhle, Asymptotic results for coalescent processes without proper frequencies and applications to the two-parameter Poisson-Dirichlet coalescent, *Stochastic Process. Appl.* **120** (2010), no. 11, 2159–2173. MR 2684740
[16] J. Pitman, Coalescents with multiple collisions, *Ann. Probab.* **27** (1999), no. 4, 1870–1902. MR 1742892 (2001h:60016)
[17] S. Sagitov, The general coalescent with asynchronous mergers of ancestral lines, *J. Appl. Probab.* **36** (1999), no. 4, 1116–1125. MR 1742154 (2001f:92019)

Goethe Universität, Robert Mayer Strasse 10, D-60325 Frankfurt am Main, Germany
E-mail address: kersting@math.uni-frankfurt.de

CIMAT, A.C., Calle Jalisco s/n, Col. Mineral de Valenciana, 36240 Guanajuato, Guanajuato, Mexico
E-mail address: jcpardo@cimat.mx

CIMAT, A.C., Calle Jalisco s/n, Col. Mineral de Valenciana, 36240 Guanajuato, Guanajuato, Mexico
E-mail address: arno@cimat.mx